ENV 790.30 - Time Series Analysis for Energy Data | Spring 2023

Assignment 4 - Due date 02/17/23

Yuxiang Ren

Directions

You should open the .rmd file corresponding to this assignment on RStudio. The file is available on our class repository on Github. And to do so you will need to fork our repository and link it to your RStudio.

Once you have the file open on your local machine the first thing you will do is rename the file such that it includes your first and last name (e.g., "LuanaLima_TSA_A04_Sp23.Rmd"). Then change "Student Name" on line 4 with your name.

Then you will start working through the assignment by **creating code and output** that answer each question. Be sure to use this assignment document. Your report should contain the answer to each question and any plots/tables you obtained (when applicable).

When you have completed the assignment, **Knit** the text and code into a single PDF file. Submit this pdf using Sakai.

R packages needed for this assignment: "xlsx" or "readxl", "ggplot2", "forecast","tseries", and "Kendall". Install these packages, if you haven't done yet. Do not forget to load them before running your script, since they are NOT default packages.\

```
# Load/install required package here
library(xlsx)
library(ggplot2)
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
library(tseries)
library(Kendall)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

Questions

Consider the same data you used for A3 from the spreadsheet "Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx". The data comes from the US Energy Information and Administration and corresponds to the December 2022 Monthly Energy Review. For this assignment you will work only with the column "Total Renewable Energy Production".

```
# Importing data set - using xlsx package
rawdata <- read.xlsx(file = "./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xls
    header = FALSE, startRow = 13, sheetIndex = 1)
read_col_names <- read.xlsx(file = "./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Sou
    header = FALSE, startRow = 11, endRow = 11, sheetIndex = 1)
colnames(rawdata) <- read_col_names
head(rawdata)
```

```
##        Month Wood Energy Production Biofuels Production
## 1 1973-01-01                129.630       Not Available
## 2 1973-02-01                117.194       Not Available
## 3 1973-03-01                129.763       Not Available
## 4 1973-04-01                125.462       Not Available
## 5 1973-05-01                129.624       Not Available
## 6 1973-06-01                125.435       Not Available
##   Total Biomass Energy Production Total Renewable Energy Production
## 1                         129.787                             403.981
## 2                         117.338                             360.900
## 3                         129.938                             400.161
## 4                         125.636                             380.470
## 5                         129.834                             392.141
## 6                         125.611                             377.232
##   Hydroelectric Power Consumption Geothermal Energy Consumption
## 1                         272.703                         1.491
## 2                         242.199                         1.363
## 3                         268.810                         1.412
## 4                         253.185                         1.649
## 5                         260.770                         1.537
## 6                         249.859                         1.763
##   Solar Energy Consumption Wind Energy Consumption Wood Energy Consumption
## 1          Not Available          Not Available                129.630
## 2          Not Available          Not Available                117.194
## 3          Not Available          Not Available                129.763
## 4          Not Available          Not Available                125.462
## 5          Not Available          Not Available                129.624
## 6          Not Available          Not Available                125.435
```

```
##   Waste Energy Consumption Biofuels Consumption
## 1                    0.157        Not Available
## 2                    0.144        Not Available
## 3                    0.176        Not Available
## 4                    0.174        Not Available
## 5                    0.210        Not Available
## 6                    0.176        Not Available
##   Total Biomass Energy Consumption Total Renewable Energy Consumption
## 1                          129.787                              403.981
## 2                          117.338                              360.900
## 3                          129.938                              400.161
## 4                          125.636                              380.470
## 5                          129.834                              392.141
## 6                          125.611                              377.232
```

```r
A04_rawdata <- data.frame(rawdata[, "Total Renewable Energy Production"])
ts_A04_rawdata <- ts(A04_rawdata, frequency = 12, start = c(1973, 1))
nrow <- nrow(A04_rawdata)
```

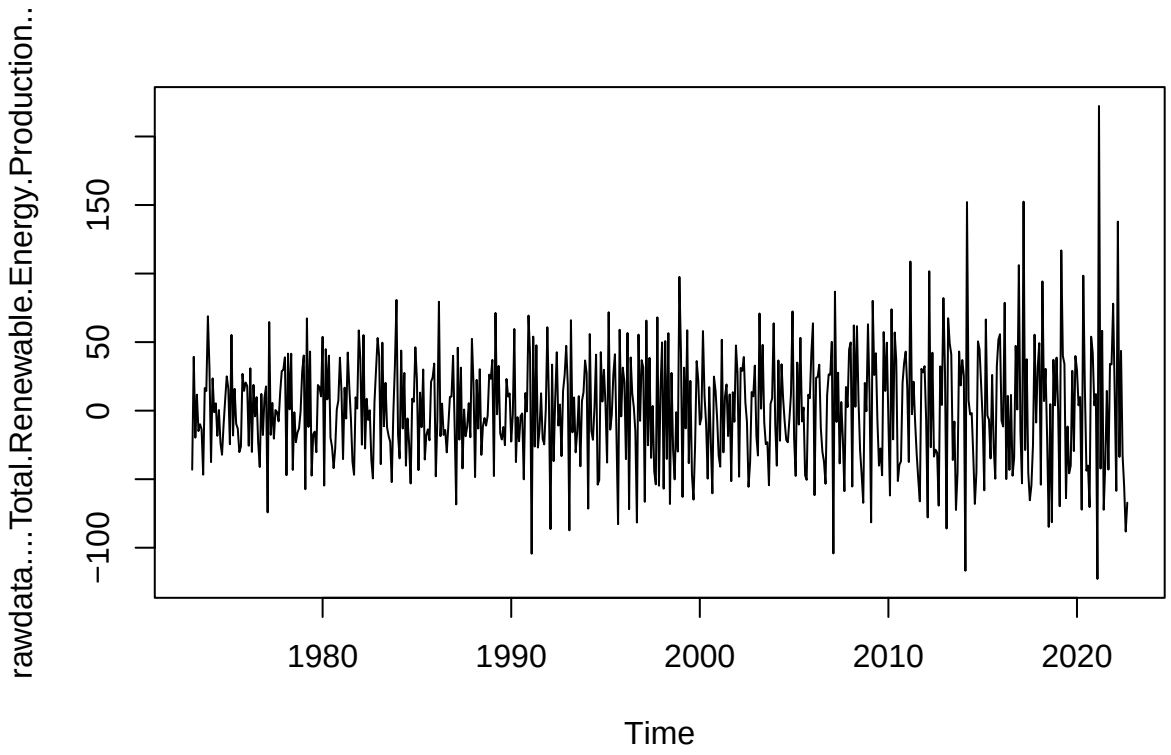
Stochastic Trend and Stationarity Tests

Q1

Difference the "Total Renewable Energy Production" series using function diff(). Function diff() is from package base and take three main arguments: * *x* vector containing values to be differenced; * *lag* integer indicating with lag to use; * *differences* integer indicating how many times series should be differenced.

Try differencing at lag 1 only once, i.e., make `lag=1` and `differences=1`. Plot the differenced series Do the series still seem to have trend?

```r
diff_1 <- diff(ts_A04_rawdata, lag = 1, differences = 1)
plot(diff_1, type = "l")
```



> Answer: It looks like there is no trend, as all values fluctuate around 0. However, it can be observed that the range of fluctuations has gradually increased with time.

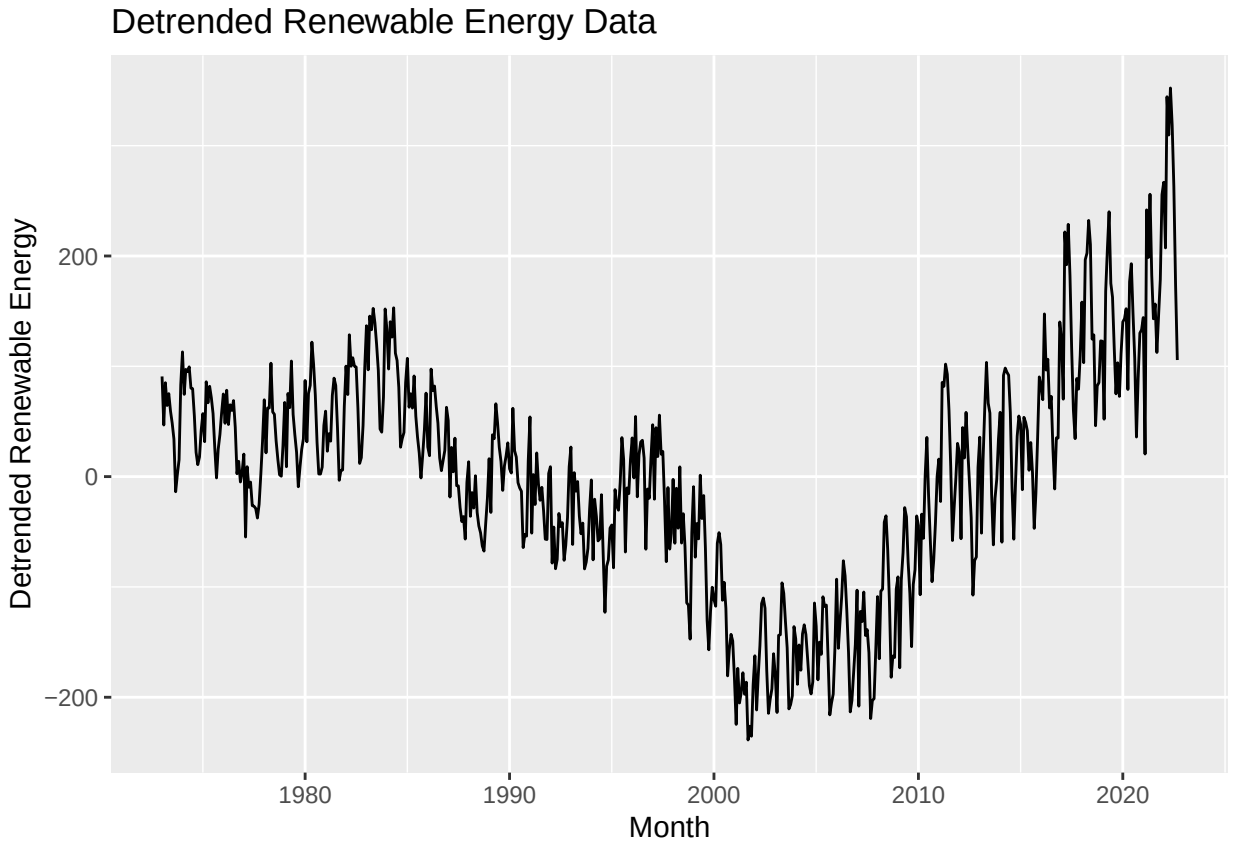
Q2

Now let's compare the differenced series with the detrended series you calculated on A3. In other words, for the "Total Renewable Energy Production" compare the differenced series from Q1 with the series you detrended in A3 using linear regression. (Hint: Just copy and paste part of your code for A3)

Copy and paste part of your code for A3 where you compute regression for Total Energy Production and the detrended Total Energy Production

```
# detrend
t <- c(1:nrow)
ren_linear_trend = lm(rawdata[, "Total Renewable Energy Production"] ~ t)
ren_beta0 = as.numeric(ren_linear_trend$coefficients[1])  #intercept
ren_beta1 = as.numeric(ren_linear_trend$coefficients[2])  #slope
ren_detrend <- rawdata$`Total Renewable Energy Production` - (ren_beta0 + ren_beta1 *
    t)
df_ren_detrend <- data.frame(rawdata$Month, ren_detrend)
names(df_ren_detrend) <- c("month", "detrend")

ggplot(data = df_ren_detrend, aes(x = month, y = detrend)) + geom_line() + xlab("Month") +
    ylab("Detrended Renewable Energy") + ggtitle("Detrended Renewable Energy Data")
```



Answer: The result of detrend in A03 is not as good as a result of the diff function. Because the detrend result in ao3 also has an obvious upward trend after 2000

Q3

Create a data frame with 4 columns: month, original series, detrended by Regression Series and differenced series. Make sure you properly name all columns. Also note that the differenced series will have only 584 rows because you loose the first observation when differencing. Therefore, you need to remove the first observations for the original series and the detrended by regression series to build the new data frame.

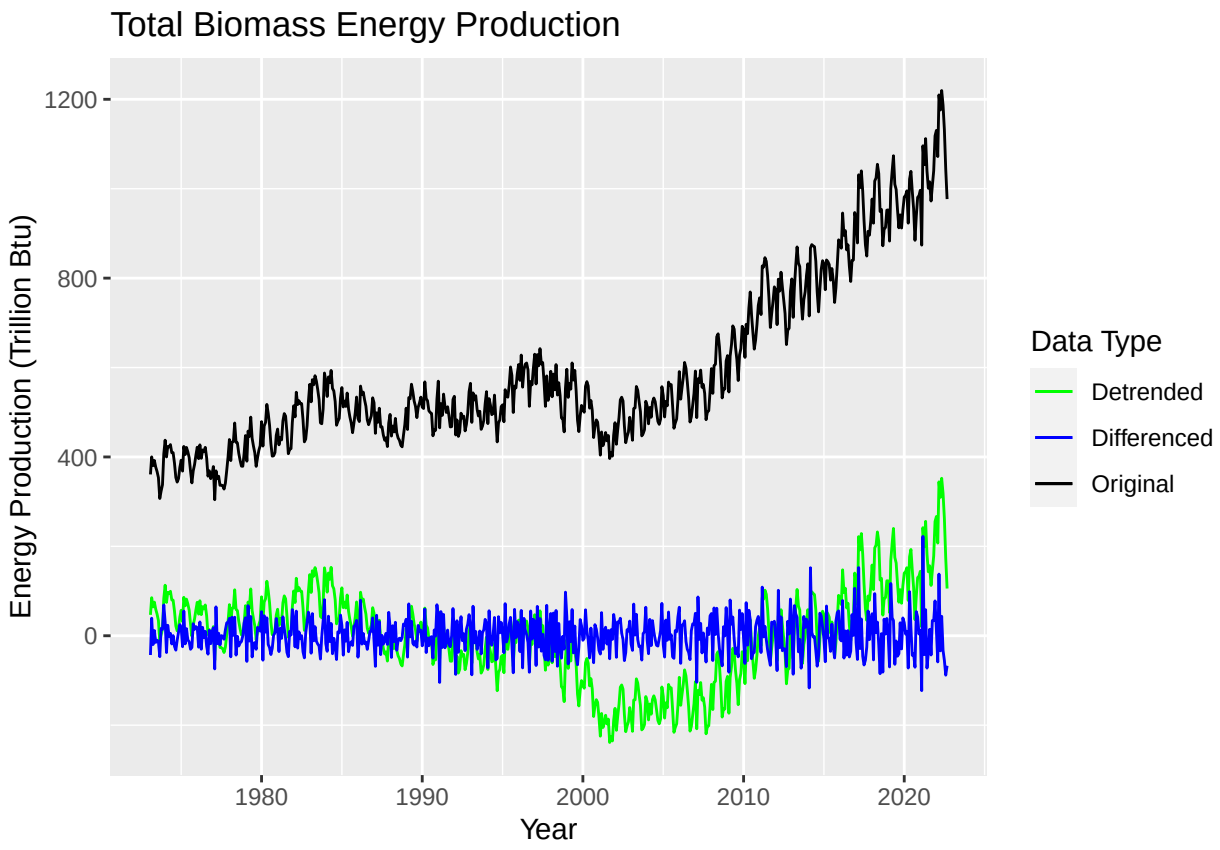
```
# Data frame not include January 1973
month596 <- rawdata[2:nrow(rawdata), "Month"]
orignial596 <- rawdata[2:nrow(rawdata), "Total Renewable Energy Production"]
detrend596 <- ren_detrend[2:597]


df <- data.frame(month596, orignial596, detrend596, diff_1)
names(df) <- c("month", "original_series", "detrended_by_regression", "differenced_series")
```

Q4

Using ggplot() create a line plot that shows the three series together. Make sure you add a legend to the plot.

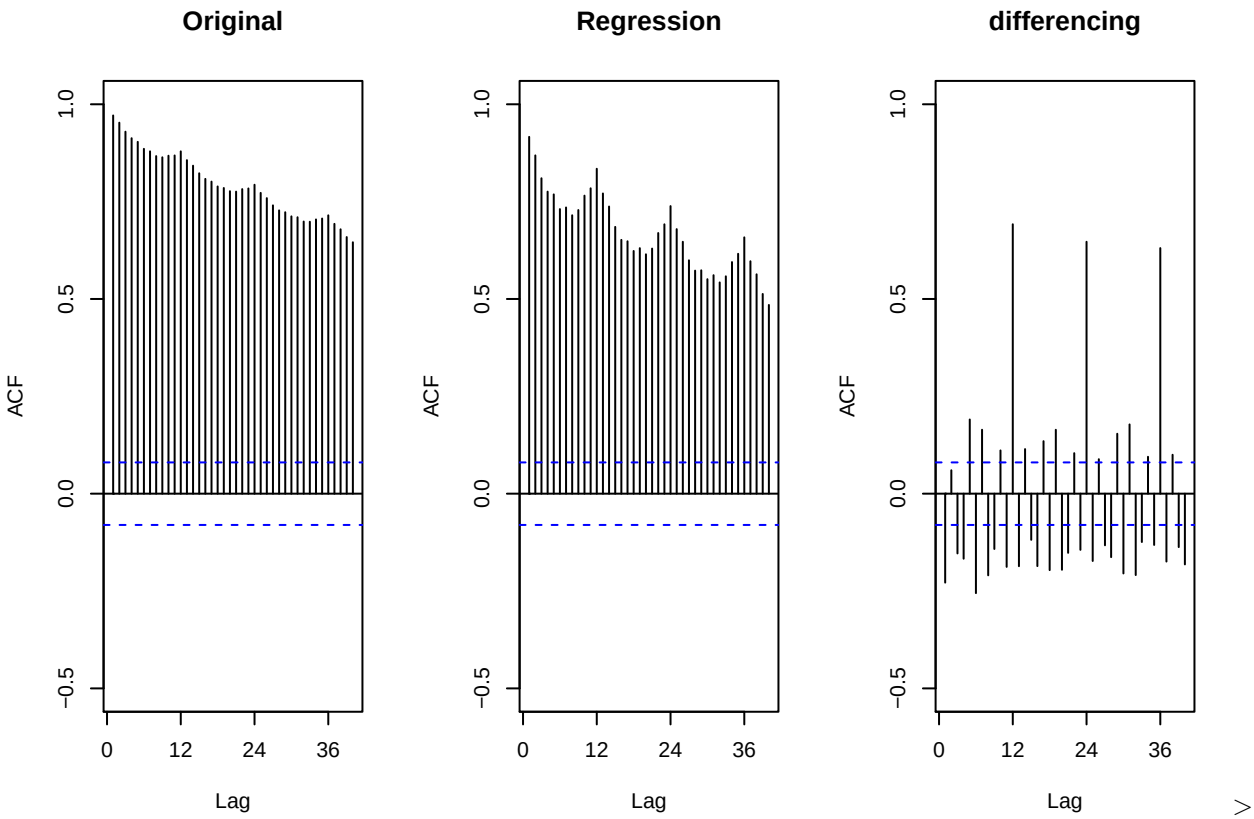
```
# Use ggplot
ggplot(df, aes(x = month, y = original_series, color = "Original")) + geom_line() +
    geom_line(data = df, aes(x = month, y = detrended_by_regression, color = "Detrended")) +
    geom_line(data = df, aes(x = month, y = differenced_series, color = "Differenced")) +
    ggtitle("Total Biomass Energy Production") + xlab("Year") + ylab("Energy Production (Trillion Btu)")
    scale_color_manual(values = c(Original = "black", Detrended = "green", Differenced = "blue"),
        labels = c("Detrended", "Differenced", "Original")) + guides(color = guide_legend(title = "Data
```



Q5

Plot the ACF for the three series and compare the plots. Add the argument `ylim=c(-0.5,1)` to the Acf() function to make sure all three y axis have the same limits. Which method do you think was more efficient in eliminating the trend? The linear regression or differencing?

```
# Compare ACFs tsdf
ts_df <- ts(df[, c(2, 3, 4)], frequency = 12, start = c(1973, 1))
par(mfrow = c(1, 3))
Acf(ts_df[, "original_series"], lag.max = 40, main = paste("Original"), ylim = c(-0.5,
    1))
Acf(ts_df[, "detrended_by_regression"], lag.max = 40, main = paste("Regression"),
    ylim = c(-0.5, 1))
Acf(ts_df[, "differenced_series"], lag.max = 40, main = paste("differencing"), ylim = c(-0.5,
    1))
```



Answer: Differencing is more efficient in eliminating the trend. Because its acf graph is no longer a simple decline, there is no obvious long-term correlation.

Q6

Compute the Seasonal Mann-Kendall and ADF Test for the original "Total Renewable Energy Production" series. Ask R to print the results. Interpret the results for both test. Whats the conclusion from the Seasonal Mann Kendall test? What's the conclusion for the ADF test? Do they match what you observed in Q2? Recall that having a unit root means the series has a stochastic trend. And when a series has stochastic trend we need to use a different procedure to remove the trend.

```
# seasonal mann-kendall
SMren <- SeasonalMannKendall(ts_A04_rawdata)
print("Results for Seasonal Mann Kendall /n")
```

```
## [1] "Results for Seasonal Mann Kendall /n"
```

```
print(summary(SMren))
```

```
## Score =  10577 , Var(Score) = 169001
## denominator =  14553
## tau = 0.727, 2-sided pvalue =< 2.22e-16
## NULL
```

```
# ADF Null hypothesis is that data has a unit root
print("Results for ADF test/n")
```

```
## [1] "Results for ADF test/n"
```

```
print(adf.test(ts_A04_rawdata, alternative = "stationary"))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_A04_rawdata
## Dickey-Fuller = -1.2055, Lag order = 8, p-value = 0.9056
## alternative hypothesis: stationary
```

> Answer: The Seasonal Mann-Kendall's result indicates that the test found a statistically significant trend in the data with a tau value of 0.727 and a two-sided p-value of less than 2.22e-16. The p-value is extremely small, indicating strong evidence against the null hypothesis of no trend in the data. The output shows that the ADF test statistic value is -1.2055, and the Lag order used in the test is 8. The p-value of the test is 0.9056, indicating that the null hypothesis contains a unit root that cannot be rejected at a 5% significance level. Therefore, the series has a stochastic trend.

Q7

Aggregate the original "Total Renewable Energy Production" series by year. You can use the same procedure we used in class. Store series in a matrix where rows represent months and columns represent years. And then take the columns mean using function colMeans(). Recall the goal is the remove the seasonal variation from the series to check for trend.

```
# remove 2022
ts_A04_rawdata_year <- ts_A04_rawdata[c(1:588), ]
years <- rawdata[c(1:588), ]

ren_data_matrix <- matrix(ts_A04_rawdata_year, byrow = FALSE, nrow = 12)
ren_data_yearly <- colMeans(ren_data_matrix)
my_year <- c(year(first(years$Month)):year(last(years$Month)))

ren_data_new_yearly <- data.frame(my_year, ren_data_yearly)
```

Q8

Apply the Mann Kendal, Spearman correlation rank test and ADF. Are the results from the test in agreement with the test results for the non-aggregated series, i.e., results for Q6?

```
# Mann Kendal
print("Results of Mann Kendall on average yearly series")
```

```
## [1] "Results of Mann Kendall on average yearly series"
```

```
print(summary(MannKendall(ren_data_yearly)))
```

```
## Score =  864 , Var(Score) = 13458.67
## denominator =  1176
## tau = 0.735, 2-sided pvalue =< 2.22e-16
## NULL
```

```r
# Spearman
print("Results from Spearman Correlation")
```

```
## [1] "Results from Spearman Correlation"
```

```r
sp_ren = cor(ren_data_new_yearly, my_year, method = "spearman")
print(sp_ren)
```

```
##                 [,1]
## my_year         1.00
## ren_data_yearly 0.87
```

```r
sp_ren_cor = cor.test(ren_data_new_yearly$ren_data_yearly, my_year, method = "spearman")
print(sp_ren_cor)
```

```
##
##  Spearman's rank correlation rho
##
## data:  ren_data_new_yearly$ren_data_yearly and my_year
## S = 2548, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##  rho
## 0.87
```

```r
# ADF.
print("Results for ADF test on yearly data/n")
```

```
## [1] "Results for ADF test on yearly data/n"
```

```r
print(adf.test(ren_data_yearly, alternative = "stationary"))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ren_data_yearly
## Dickey-Fuller = -0.68508, Lag order = 3, p-value = 0.9654
## alternative hypothesis: stationary
```

Answer: The result of the yearly Mann-Kendall test is similar to the seasonal test, which found a statistically significant trend in data with a tau value of 0.735 and a two-sided p-value of less than 2.22e-16. The result for the yearly ADF test is also similar to the test based on seasonal data, which indicates a stochastic trend due to the p-value being 0.9654. From the Spearman correlation rank test, the output shows a strong positive correlation between the yearly renewable energy data and the year variable, with a correlation coefficient of 0.87 and a p-value less than 2.2e-16.